PROCEEDINGS OF THE 1984 INTERNATIONAL CONFERENCE ON OCCUPATIONAL ERGONOMICS

USE OF VALIDITY MEASURES IN THE SELECTION OF PHYSICAL SCREENING TESTS

J.W. Nottrodt and E.J. Celentano

Human Engineering Section
Defence and Civil Institute of Environmental Medicine
Downsview, Ontario, Canada
M3M 3B9

ABSTRACT

The purpose of this paper is to demonstrate how measures of validity can be used in the selection of screening tests. Both logical and statistical validity approaches are discussed. The analysis dealt primarily with a statistical method which assessed the usefulness of a test cutoff score in predicting task performance. A numerical example is included to illustrate the steps which could be used to maximize the confidence in the selection of a valid test.

INTRODUCTION

In order to better utilize personnel and ensure their safety and human rights, there has been a growing need for screening tests for physically demanding occupations in both the military and industrial sectors. Many factors are involved in deciding which tests would be appropriate as screening tools to aid in personnel management. These factors include test reliability, safety, cost and time constraints, but most important are the measures of validity between the actual occupational demands and the tests used to assess an individual's capabilities.

Validity Approaches

Two types of approaches can be used to establish the validity of a screening test. Both involve comparisons against some selected criteria, in this case the occupational requirements which are physically demanding. The first type are logical approaches and involve the use of content- and construct-referenced strategies to identify valid relationships between tests that measure physical capability and tasks that represent occupational physical requirements (Arnold et al., 1982). An occupational analysis to isolate the physically-limiting task (or group of tasks) is required. This task itself (content-referenced) or elements of it (construct-referenced) can then be used in the screening test.

The second type are statistical approaches which primarily attempt to identify predictive relationships between test and task performances. An example of this method involves the use of correlational techniques (Norman et al., 1983; Reilly et al., 1979). The "r" values of a number of tests could be compared to determine which test has the highest prediction power. If high correlations are not demonstrated, it has been concluded that the tests are questionable as valid predictors of task performance (Norman et al., 1983). The main difficulty with this method is in deciding what criteria define a high "r" value.

Another example of a statistical method determines the usefulness of a performance standard or cutoff score on a screening test in correctly sorting individuals into success and failure groups based on task performance (Berk, 1976; Douglass and Safrit, 1983; Washburn and Safrit, 1982). This method utilizes four analytical procedures which can be used first to determine the optimal performance standards on a number of candidate screening tests and then used to determine which test is most useful as a predictor of task performance. Both of the statistical methods require cross validation samples to establish concurrent validity (Berk, 1976; Ferguson, 1976; Reilly et al., 1979; Washburn and Safrit, 1982).

PURPOSE

The purpose of this paper is to show how these validity approaches can be employed in selecting the most appropriate screening test. The validity approaches involving the construct-referenced strategy and the correlational techniques, were used to identify screening tests which possessed potentially useful predictive relationships with task performance and were discussed briefly. The focus was primarily on the statistical approach whi h assesses test usefulness since it can be used to determine not only which test is the most useful predictor, but also the optimal performance standard on that particular test. A numerical example was obtained from data collected in a previous study involving 56 subjects (31 males and 25 females), who performed a task, identified as a limiting occupational requirement, and a number of candidate screening tests (unpublished data).

ANALYSIS AND DISCUSSION

The task identified as the limiting occupational requirement involved the lifting of a box (dimensions 41x60x25 centimeters) onto the flatbed of a vehicle. The maximum lifting capacity of the subjects for this task was determined using an incremental protocol with a starting level at no less than 7 kilograms to end at maximum capability or 91 kilograms. The criterion for a successful performance of the task, as determined from an analysis of the task

characteristics, was the single lift of a 32 kilogram box to a flatbed height of 133 centimeters. The task performance scores for males and females appear in Table 1.

Using a construct-referenced strategy, the following candidate screening tests were selected:

1) a dynamic incremental lift to 183 centimeters starting at 18 kilograms to maximum capability or 91 kilograms;

2) a maximal static exertion at a height of 110 centimeters; and

3) a measurement of body mass.

Because the action of lifting was common, construct validity was high between the task and the two strength tests. This argument was especially true for the task and the dynamic lift test as both involved similar protocols. Although it involved a weaker argument construct validity, the measurement of body was selected because the lift action of the t involved the attribute of strength and there evidence in the literature of a relationship between various strength tests and body mass (Elbel, 1949; Laubach, 1976; Watson and O'Donovan, 1977). The test performance scores for males and females are summarized in Table 1.

TABLE 1

Performance Data on the Task and the Screening Tests.

Parameter (kg)	Male (n=31) $\bar{X}$	s.d.	Female (n=25) $\bar{X}$	s.d.
Incremental Box Lift Task	43.1	8.1	22.4	5.7
Incremental Lift Test to 183 cm	43.9	7.3	22.9	4.5
Maximal Static Exertion at 110 cm	46.3	9.1	24.3	5.6
Measurement of Body Mass	75.0	9.6	62.2	9.4

Correlations between performance scores on the lift task and the three tests were calculated (Table 2). All the correlations were significant at the p(0.01) level. The "r" values for the males and females ranged from 0.55 to 0.72. Norman et al. (1983) suggested that the validity of predictive relationships with "r" values in this range was questionable because the shared variances (r^2) between the task and test performance scores were too low. If an argument based upon correlational validity alone was used in identifying valid relationships between performance of the task and the three predictors, it would appear that these tests would not be appropriate.

TABLE 2

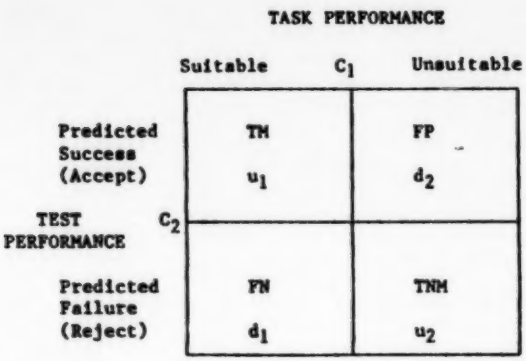
Correlations Between Performance on the Task and the Screening Tests.

Test	Correlation (r) Male	Female
Incremental Lift Test to 183 cm	0.71	0.69
Maximal Static Exertion at 110 cm	0.66	0.60
Measurement of Body Mass	0.55	0.72

N.B. All correlations significant at p(0.01) level.

An alternative argument of test validity involved the assessment of the usefulness of performance standards (i.e., cutoff scores) on screening tests in correctly sorting subjects into success and failure groups based on their task performances. This method utilized four analytical procedures as measures of validity: 1) classification of outcome probalities; 2) correlational approach; 3) utility analysis; and 4) incremental validity (Berk, 1976). Procedures 1, 2 and 3 were used to determine the optimal cutoff scores on the screening tests. The fourth procedure, incremental validity, compared the probabilities of correct decisions between the different tests.

The classification of outcome probabilities was determined by calculating the correct number of decisions made when sorting individuals into success and failure performance groups. The correlational approach, which used the calculation of correct decisions of the successful group, also accounted for the probability of making false positive and false negative decisions. The utility analysis involved the calculation of a maximal utility coefficient from measures of utility and disutility. Within the measures of utility and disutility, the effects of emphasizing true success, true failure, false positive or false negative probabilities were considered. Two estimates of incremental validity were calculated for each test and were completed by comparing the validity coefficient for the classification of outcome probabilities of one test with: 1) a base rate value; and 2) the same validity coefficient of another test. The classification diagram in Figure 1 formed the basis for the calculations of the validity coefficients of the four analytical procedures (the formulae are summarized in Table 3). Whereas one procedure may not provide sufficient evidence, the combination of the four procedures, when yielding similar results, provide the basis for a more confident selection of both a valid test and performance standard (Douglass and Safrit, 1983).



where: C_1 = cutoff score denoting suitable/unsuitable task performance

C_2 = cutoff score denoting predicted success/failure test performance

FN = false negative scores

FP = false positive scores

TM = true success scores

TNM = true failure scores

d_1, d_2, u_1, u_2 = coefficients used in utility analysis

Figure 1. Classification Diagram (adapted from Washburn and Safrit, 1982).

TABLE 3

Validity Coefficient Formulae for the Four Analytical Procedures.

1) Classification of Outcome Probabilities (C)

$C = p(TM) + p(TNM)$

2) Classification Approach (ϕ_{vc})

$$\phi_{vc} = \frac{TM - BR(SR)}{\sqrt{BR(1 - BR)SR(1 - SR)}}$$

where: $TM = p(TM)$
$BR = p(TM) + p(FN)$
$SR = p(TM) + p(FP)$

3) Utility Analysis (involves calculation of 3 coefficients)

i) Utility (U)

$U = p(TM)u_1 + p(TNM)u_2$, where u_1 and $u_2 > 0$

ii) Disutility (D)

$D = p(FN)d_1 + p(FP)d_2$, where d_1 and $d_2 < 0$

iii) Maximum Utility (γ)

$\gamma = (U + D)N$, where N = number of subjects

4) Incremental Validity (Ω)

i) comparison between using test and not using test (base rate)

$\Omega = [p(TM) + p(TNM)] - BR$, where $BR = p(TM) + p(FN)$

ii) comparison between two tests

$\Omega = [p_1(TM) + p_1(TNM)] - [p_2(TM) + p_2(TNM)]$.

The male and female data were combined in the analysis since the validity estimates do not require a unimodal distribution of test scores (Douglass and Safrit, 1983). To identify the optimal test performance standards, three possible cutoff scores were analyzed for each test and their validity coefficients are summarized in Table 4. Using the first three analytical procedures, it was determined that a performance standard of 32 kilograms provided the optimal probability of correct decisions on both strength tests. The results of the anlaysis to determine the optimal cutoff score on the body mass test were not as conclusive. A performance standard of 70 kilograms body mass yielded the best results. The optimal performance standards on the dynamic and static strength tests showed a similar number of correct sorting decisions (classification of outcome probabilities X 100) of individuals into success and failure performance groups (95% and 93%, respectively). Both estimates of incremental validity, i.e., the comparisons to either a base rate value or another test, showed that the strength tests provided more correct decisions than the measurement of body mass (Table 5). Therefore it was concluded that the strength tests would be more appropriate choices as screening tests.

TABLE 4

Summary of Validity Coefficients to Identify Optimal Performance Standards.

Cutoff Score on Task Performance = 32 Kilograms

Test	Cutoff Score	C	ϕ_{yc}	Ya	Yb	Yc	Yd
Incr. Lift Test to 183 cm	27 kg	0.82	0.68	25.8	35.8	56.6	51.0
	32	0.95	0.90	47.6	48.7	77.8	71.7
	36	0.89	0.80	42.6	38.6	68.3	62.7
Max. Static Exertion at 110 cm	27	0.83	0.62	27.4	35.3	57.7	51.5
	32	0.93	0.85	45.4	45.4	77.3	68.3
	36	0.93	0.84	51.4	45.9	76.3	66.6
Measurement of Body Mass	60	0.70	0.70	24.6	30.2	54.9	39.2
	70	0.79	0.60	28.6	24.6	51.5	43.7
	80	0.61	0.35	11.8	-9.5	21.8	14.6

N.B. For maximal utility (γ), the weightings for the utility and disutility coefficient calculations (see Table 3) were:

a— $d_1/d_2 = -1/-2$, $u_1/u_2 = 1/1$

b— $d_1/d_2 = -2/-1$, $u_1/u_2 = 1/1$

c— $d_1/d_2 = -1/-2$, $u_1/u_2 = 2/1$

d— $d_1/d_2 = -2/-1$, $u_1/u_2 = 1/2$

TABLE 5

Incremental Validity Coefficients (Ω).

	Versus Base Rate	Versus Other Test T1	T2	T3
1) Incr. Lift Test to 183 cm	0.39	—	0.02	0.16
2) Max. Static Exertion at 110 cm	0.37	—	—	0.14
3) Measurement of Body Mass	0.24	—	—	—

CONCLUSIONS

This paper has demonstrated that logical and statistical approaches to establishing validity can be used to select an appropriate screening test. The following steps can be employed to maximize the confidence in the selection of a valid test:

1) use content- and construct-referenced strategies to identify potential tests which have predictive relationships with task performance;

2) use correlational statistics to identify the tests which have a significant predictive relationship with task performance; and

3) use the combination of four analytical procedures (classification of outcome probabilities, correlational approach, utility analysis, incremental validity) to determine the optimal performance standards of the tests, and then to determine which test is the most useful predictor.

Cross validation with samples of sufficient size are necessary to ensure that:

1) a significant relationship exists between the selected screening test and the task;

2) the performance standard on the screening test is the optimal cutoff score; and

3) the screening test is the most useful predictor of task performance.

It has been suggested that samples should contain at least 100 subjects in each of the true success and true failure cells (refer to Figure 1) (Berk, 1976; Washburn and Safrit, 1982).

If the decision between two or more tests cannot be settled clearly by the analysis to determine the most useful predictor, as was the case with the two strength tests in the example, other factors may need to be assessed. These include test safety, reliability and practicality in terms of cost, time consumption, administration, data reduction and interpretation.

REFERENCES

Arnold, J.D., Rauschenberger, J.M., Soubel, W.G. and Guion, R.M. (1982) Validation and Utility of a Strength Test for Selecting Steelworkers. Journal of Applied Psychology, 67(5): 588-604.

Berk, R.A. (1976) Determination of Optimal Cutting Scores in Criterion-referenced Measurement. Journal of Experimental Education, 45: 4-9.

Douglass, J.A. and Safrit, M.J. (1983) An Empirical Approach to the Validation of a Criterion Referenced Measure of Motor Performance. Journal of Human Movement Studies, 9: 57-69.

Elbel, E.R. (1949) Relationship Between Leg Strength, Leg Endurance and Other Body Measurements. Journal of Applied Physiology, 2: 197-207.

Ferguson, G.A. (1976) Statistical Analysis in Psychology and Education. McGraw-Hill, New York, 1976.

Laubach, L.L. (1976) Muscular Strength of Women and Men: A Comparative Study. Aerospace Medical Research Lab, Wright-Patterson AFB, AMRL-TR-75-32.

Norman, R., Sharratt, M., Sutherland, D. and Baleshta, M. (1983) The Question of Validity of Isometric Strength Tests as a Pre-employment Screening tool. In Proceedings of the Human Factors Association of Canada, 16th Annual Meeting, 1983, 15-18.

Reilly, R.R., Zedeck, S. and Tenopyr, M.L. (1979) Validity and Fairness of Physical Ability Tests for Predicting Performance in Craft Jobs. Journal of Applied Psychology, 64(3): 262-274.

Washburn, R.A. and Safrit, M.J. (1982) Physical Performance Tests in Job Selection: A Model for Empirical Validation. Research Quarterly for Exercise and Sport, 53(3): 267-270.

Watson, A.W.S. and O'Donovan, D.J. (1977) Factors Relating to the Strength of Male Adolescents. Journal of Applied Physiology, 43(5): 834-838.